



**University of
Zurich**^{UZH}

**Zurich Open Repository and
Archive**

University of Zurich
University Library
Strickhofstrasse 39
CH-8057 Zurich
www.zora.uzh.ch

Year: 2018

Are ratings the worst form of credit assessment apart from all the others?

Bloechlinger, Andreas ; Leippold, Markus

Abstract: We present a prediction model to forecast corporate defaults. In a theoretical model, under incomplete information in a market with publicly traded equity, we show that our approach must outperform ratings, Altman's Z-score, and Merton's distance to default. We reconcile the statistical and structural approaches under a common framework, i.e., our approach nests Altman's and Merton's approaches as special cases. Empirically, we cannot reject the superiority of our approach. Furthermore, the numbers of observed defaults align well with the estimated probabilities. Finally, with rank transforms, we obtain cycle-adjusted forecasts that still outperform ratings.

DOI: <https://doi.org/10.1017/S0022109017000874>

Posted at the Zurich Open Repository and Archive, University of Zurich

ZORA URL: <https://doi.org/10.5167/uzh-127427>

Journal Article

Accepted Version

Originally published at:

Bloechlinger, Andreas; Leippold, Markus (2018). Are ratings the worst form of credit assessment apart from all the others? *Journal of Financial and Quantitative Analysis*, 53(1):299-334.

DOI: <https://doi.org/10.1017/S0022109017000874>

Are Ratings the Worst Form of Credit Assessment Apart from All the Others?*

Andreas Blöchlinger[†]

Zürcher Kantonalbank, University of St.Gallen

Markus Leippold[‡]

University of Zurich

Basile Maire[§]

Zürcher Kantonalbank

First Version: July, 2011

Current Version: July 25, 2012

*The content of this paper reflects the personal view of the authors. In particular, it does not necessarily represent the opinion of Zürcher Kantonalbank. Markus Leippold gratefully acknowledges financial support from the Swiss Finance Institute (SFI) and Bank Vontobel.

[†]**Correspondence Information:** Andreas Blöchlinger, Zürcher Kantonalbank, Josefstrasse 222, CH-8010 Zurich, Switzerland; tel: +41 44 292 45 80, <mailto:andreas.bloechlinger@zkb.ch>, andreas.bloechlinger@unisg.ch

[‡]**Correspondence Information:** Markus Leippold, University of Zurich, Plattenstr. 22, CH-8032 Zurich, Switzerland; tel: +41 44 634 50 69, <mailto:markus.leippold@bf.uzh.ch>

[§]**Correspondence Information:** Basile Maire, Zürcher Kantonalbank, Josefstrasse 222, CH-8010 Zurich, Switzerland; tel: +41 44 292 45 79, <mailto:basile.maire@zkb.ch>

Are Ratings the Worst Form of Credit Assessment Apart from All the Others?

Abstract

Ending the dependence on rating agencies is a top priority for the Financial Stability Board, which coordinates the G20 financial policies. Rating agencies have been accused of contributing to the recent financial crisis by misjudging the creditworthiness of mortgage-backed securities. Their downgrading practice of sovereigns and corporations is said to amplify procyclicality and to fuel market uncertainty. However, it is one thing to criticize rating agencies, quite another to find an appropriate alternative. For the case of corporate issuer ratings, we show that there is at least one such alternative. From a statistical viewpoint, it outperforms credit ratings. From a practical viewpoint, it is straightforward to implement and depends solely on publicly available data. Therefore, we conclude that for the case of corporate credit risk, appropriate alternatives to credit ratings do exist!

JEL Classification Codes: G01, G18, G24

Key Words: Credit Rating Agencies, Distance-To-Default, Z-Score, Zeta-Score, Default Prediction

1 Introduction

The former chief economic advisor of the Obama administration once set up a parable that compared finance theorists with “ketchup economists”:

“They have shown that two quart bottles of ketchup invariably sell for twice as much as
one quart bottles of ketchup”Summers (1985)

Ketchup economists conclude from this finding that the ketchup market is perfectly efficient. Like parallels apply to credit ratings. Rating agencies assign characters to firms and sell the information about these allocations. The more characters, and the earlier in the alphabet, the more likely is the firm perceived to pay back its debt. Similar to the ketchup economists, rating agencies provide us with relative assessments. Though investors must use ratings as an absolute indication of the likelihood of receiving the money owed to them in accordance with the terms on which they invested. Furthermore, agencies keep their ratings stable in times of crisis as long as the relative rank order remains balanced, even if the firms’ default probabilities increase. Rating agencies thus do not provide us with an answer to what seems an important question of which rating is assigned to what level of default probability. We call this the *level* question. Despite the agencies’ focus on relative performance, ratings do not provide us information about the marginal increase in default risk between the ratings. How much more likely is a firm to fail if it is downgraded from *AA* to *A*? We call this the *shape* question. What the agencies seem to provide is discriminatory power, that is, the ability to distinguish between surviving and defaulting borrowers. With the recent crisis, critics have not been slow to note that the agencies showed a weak performance, even regarding the discriminatory power of their rating assignments. Yet alternatives are hard to find. Hilscher and Wilson (2012) provide a simple model that outperforms the agencies’ ratings in terms of discriminatory power, and that better explains the fluctuations in the credit markets. However, the two central questions whether the new contender explains the absolute level of default risk, and whether the difference in predicted default risk between two borrowers agrees with the real increase in default risk – questions regulators should be concerned about – remain unanswered.

In this paper, we present a simple credit risk model that feeds only on publicly available financial data. Our model is transparent, replicable, and easy to implement. In a carefully executed out-of-sample analysis, we provide empirical evidence that the model we propose significantly outperforms the default forecasts implied by credit ratings. Most importantly, our model answers the *level*, and *shape* questions. Hence, our model may serve as an appropriate alternative to credit ratings from both a regulatory and an economic perspective.

Critics blame the agencies, in addition to their poor performance, for amplifying procyclicality and increasing overall market uncertainty. They are accused to do so not only in the case of corporate ratings, but also in the case of sovereign debt. When S&P unprecedentedly downgraded the United

States' long-term debt from its triple-A rating in August 2011, it prompted a global sell-off and a spike in market volatility not seen since December 2008. But also European countries have been on the agencies' radar. Since spring 2010, the rating agencies relegated Greece, Portugal, and Ireland to "junk" status, a move that many EU officials say has accelerated a burgeoning eurozone sovereign debt crisis. In January 2012, amid continued eurozone instability, S&P downgraded nine eurozone countries, stripping France and Austria of their triple-A ratings. Whether these downgrades are justified or not, fact is that in most cases the agencies' action has led to disruptions in financial markets and fueled investors' fear about the vulnerability of the financial stability.

Despite the harsh criticism, credit ratings from the big three rating agencies remain the most common and widely used measure of credit quality. Alternative approaches have been slow to emerge and most firms and regulators continue to use their ratings. This sustained power of the big agencies stems from their designation as government-approved rating firms.¹ In 1975, the U.S. Securities and Exchange Commission (SEC) began choosing, which raters could be used to determine the minimum capital requirements for financial institutions to trade certain debt securities. The three raters initially chosen, S&P, Moody's, and Fitch, were christened as "nationally recognized statistical rating organizations" (NRSROs). Over the years, the SEC added more rating agencies to the list of NRSROs. However, the big three rating agencies maintained their dominant positions and their ratings have enjoyed the status of a quasi-regulatory standard and a commonly agreed principle heavily hardwired into defining capital standards, pricing corporate bonds, investment processes, and contract specifications of structured products.

With the recent financial crisis, however, public pressure on the rating agencies' hegemony is mounting. The Financial Stability Board, coordinating the G20 financial policies, has asked governments and regulators to find ways to break the dependency on credit ratings. Both the United States and Europe have taken steps to regulate the three main rating agencies and ensure more transparency and competitiveness. In July 2010, the United States enacted the Dodd-Frank Wall Street Reform and Consumer Protection Act (Dodd-Frank Act),² which created an Office of Credit Ratings at the Securities and Exchange Commission to hold rating agencies accountable and protect investors and businesses. This commission has to review any regulation that makes any references to credit ratings. In a subsequent step, the commission is supposed to remove any requirement of reliance on credit ratings and to substitute these with an alternative approach the commission finds appropriate. Also in Europe moves are underway to regulate the credit ratings agencies more tightly. In early 2011, the EU established an independent authority, the European Securities and Markets Authority (ESMA), tasked with regulating the activities of rating agencies relative to EU standards. The ESMA initiative basically parallels the Dodd-Frank Act. However, for the time being, the markets – and some of the world's biggest economies – remain at the rating agencies'

¹For an interesting historical review of how the judgments of creditworthiness by the rating agencies attained the force of law, we refer to Partnoy (1999).

²See <http://www.sec.gov/rules/proposed/2011/34-64352.pdf>.

mercy. Indeed, it is one thing to identify the shortcomings of credit ratings, quite another to find an appropriate alternative standard. Despite all the criticism, the question remains unanswered as to what an appropriate alternative to credit ratings might be. Nevertheless, we might agree on some common principles, which we briefly sketch below.

Firstly, the probability of default should be calibrated, that is, it should answer the shape and the level questions. A calibrated model matches predicted and observed default rates across the entire span of the data. Unfortunately, the design of an appropriate test statistic for validating the forecasting power along these two dimensions is very subtle, as credit default data is sparse and not i.i.d. distributed. Some suggest replacing corporate and sovereign ratings with credit-default-swap prices. However, these prices are pro-cyclical and incorporate a systematic risk premium which induces a bias in the default forecast.³ Furthermore, two obligors with equal default probabilities can have different credit-default-swap spreads, depending on the economic state in which they are expected to fail, see Coval et al. (2009). Hence credit spreads do not answer the shape and the level questions either and are thus no viable alternative to agency ratings.

Secondly, an appropriate alternative can only be one with a discriminating power at least as good as that of credit ratings. Most importantly, alternatives must perform reasonably well in times of market distress.

Thirdly, an appropriate alternative should be transparent, especially so if used for regulatory purposes. The joint notice issued by the Board of Governors of the Federal Reserve System, the Federal Deposit Insurance Corporation, and the Office of the Comptroller of the Currency requires that a potential alternative must be sufficiently transparent, unbiased, replicable, and defined to allow banking organizations of varying size and complexity to arrive at the same assessment of creditworthiness for similar exposures and to allow for appropriate supervisory review.⁴ In contrast, the credit ratings of S&P, Moody's, and Fitch all are black boxes as they are all based on proprietary models and data.⁵

Fourthly, we believe that credit ratings should preferably be point-in-time and not through-the-cycle forecasts.⁶ To justify our statement, we refer to the ongoing debate on bank regulation. It must be in the regulator's interest to link regulatory capital requirements as closely as possible to the banks' economic capital.⁷ With more transparent credit risk buffering, taking into account the increased default risk in times of recession, credit ratings should naturally lead to a cyclical behavior of capital requirements. Rating agencies such as S&P disregard cyclical effects. Credit ratings do not anticipate the higher bankruptcy risk during recessions. However, a powerful default predictor

³See, e.g., Blöchliger (2010).

⁴See Federal Register Volume 76, Number 245 (Wednesday, December 21, 2011), <http://www.gpo.gov/fdsys/pkg/FR-2011-12-21/html/2011-32073.htm>.

⁵Indeed, any rating provided by a rating company may be flawed with respect to replicability, as these companies have to make a living out of it.

⁶There are different views on this subject. See, e.g., Altman and Rijken (2006).

⁷See, e.g., Crouhy et al. (2001).

must anticipate the increased default risk even if the heightened default intensity is expected to be only transitory and not permanent. The default intensity may be expected to increase temporarily during economic downturns, but may decrease again in an economic upswing if the borrower indeed survives this transition period. This view is also implicitly supported by the Dodd-Frank Act.⁸

From a stability viewpoint, we may finally argue that credit quality should be expressed as a continuous variable. When an ordinal credit rating changes, the implied default probability changes discontinuously. If a change in rating hits a company whose failure poses a systemic risk, this discontinuity may well spread over into the stock market leading to a substantial drop in the company's share price. Hence, discontinuous changes during difficult times may further nourish market uncertainty by giving incentive for creditors to run, precipitating a collapse and contagion. A prominent example of such an epidemic effect is the downgrade of AIG in 2008 leading to multiple collateral calls, increased liquidity stress, and falling market confidence.

We note that the above list is by no means exhaustive. Nevertheless, it represents a list of minimal requirements for potential candidates, which may serve as appropriate alternatives to credit ratings. Furthermore, it allows us to position our contribution along these principles: By using the appropriate test statistics which overcome the problem of sparseness and dependency in default data, our out-of-sample study shows that, with a simple credit risk model, we can significantly outperform the forecasting power of credit ratings. Since we rely only on publicly available financial data and our credit model is straightforward to implement, we also fulfill the requirement of transparency and replicability. With the calibration tests of Blöchlinger and Leippold (2011) for the default probabilities, we ensure unbiased estimates of default probabilities. Furthermore, we provide two versions of our model, one that generates a point-in-time forecast and that generates a through-the-cycle default probability forecast. We find that our through-the-cycle forecast performs at least as good as the forecast based on credit ratings. For our point-in-time forecast, we find that it significantly outperforms the forecast based on credit ratings. Hence, we contribute to the literature on credit risk by showing that a model does exist, which is in line with the principles outlined above and which can be deemed an appropriate alternative to corporate credit ratings.

The remainder of the paper is structured as follows. In Section 2, we present the literature most closely related to our work. In Section 3, we introduce the market-based and accounting-based methodologies we use for the design of our credit risk model. In Section 4, we describe the data underlying the calibration of our model and the out-of-sample analysis to study the forecasting ability of different model specifications and credit ratings. In Section 5, we present our statistical analysis. We split our analysis into an in-sample and an out-of-sample analysis. Section 6 concludes.

⁸See, e.g., the Amendments to Rule 15c3-1, <http://www.sec.gov/rules/proposed/2011/34-64352.pdf>.

2 Related Literature

Predicting corporate failure has been a long standing problem in credit risk research. There is an abundant literature on this topic. Therefore, we will mention only those papers that are directly related to our work. One of the most influential and early contribution is the Altman (1968) Z-score model that employs information from accounting ratios to form a forecast of default. The Altman's Z-score relies mostly on information obtained from companies' financial statement. The Z-score model was subsequently refined in Altman et al. (1977) to the so called Zeta-score model. We use this model for our analysis of accounting-based default forecasts.

The basic model to forecast defaults based on market data is Merton's (1974) bond pricing (or distance-to-default) model. Instead of using accounting data, the main input for this model is the value of equity and its volatility to obtain the firm's default probability. Hence, the model does use the market value of debt rather than the book value of debt, as accounting models do.⁹ For our purpose, we implement the version of the distance-to-default model as described in Bharath and Shumway (2008). Their focus is on exploring the performance of this model compared to simpler alternatives. Indeed, they find that the simpler model performs slightly better than the distance-to-default model in out-of-sample default forecasts. However, they do not compare the forecasting performance to the performance of the agencies' credit ratings.

Shumway (2001) develops a hazard model for default forecasting combining both accounting and market data. He argues that well-established bankruptcy prediction models such as, e.g., the Altman (1968) Z-score model, are not correctly specified as they do not consider all the available firm-year observations. In contrast, a hazard model takes into account all the available observations for the bankrupt and non-bankrupt firms. Shumway (2001) shows that the inclusion of these additional observations produces efficient and consistent estimates of the coefficients. Furthermore, he documents that using a discrete hazard model, many of the accounting ratios incorporated in accounting-based models are not statistically significant for predicting bankruptcy. Combining both accounting ratios and market-driven variables into a hazard model outperforms the purely accounting-based models in out of sample forecasts. In our paper, we follow Shumway's (2001) suggestion in that we also construct a hybrid model based on accounting and market data.

A rich literature on the role of credit ratings has emanated in the wake of the US corporate credit meltdown in 2001 and 2002. An early review of this literature is given in Cantor (2004). Later literature includes, e.g., Boot et al. (2006), where they theoretically study two important functions of credit rating agencies, the dissemination of ratings information to market participants and the facilitation of financial regulation and contracting. In an empirical study, Becker and Milbourn (2011) find a significantly negative link between competition and quality. Hence, calls for more competition in the ratings industry¹⁰ may have counter-productive effects, since increasing competition may

⁹Vassalou and Xing (2004) use the Merton model to explore the effect of default risk on equity return.

¹⁰An increase in competition was required, e.g., by the US Department of Justice (1998).

impair the reputational mechanism that seemingly underlies the provision of good quality ratings.¹¹ With the recent financial crisis, a strand of literature emerged that particularly called into question the rating practice for structured products.¹²

Closely related to our study is the paper by Hilscher and Wilson (2012). They find that credit ratings are dominated as predictors of corporate default by a simple model based on publicly available financial information. They conclude that ratings are poor measures of raw default probabilities. Hilscher and Wilson (2012) do not apply calibration tests, that is, they do not answer the shape and the level questions. Furthermore, they estimate their model using data up to 2003 and evaluate their model from 2004 to 2008. Hence, the interesting period over the recent financial crises is left out. They conclude that, for the out-of-sample period, the relative difference between their default forecasts and the forecasts based on credit ratings is larger than for their full sample. However, they do not provide any thorough statistical analysis of their out-of-sample forecasts. Furthermore, to test the cumulative forecast accuracy, as opposed to the marginal forecast accuracy, Hilscher and Wilson (2012) take accounting data as of January which, however, is not available to the rating agencies at this point in time. Therefore, the comparison of the forecasting performance starts with giving a strong disadvantage to the rating agencies. Finally, their definition of default is taken from the KRIS data set and not from the S&P's 'D'-rating. This raises the question whether these two definitions of the default state differ. If they do, the comparison will become biased.

There is overwhelming evidence that defaults are dependent and tend to cluster. The issue of default clustering has been studied, e.g., in Das et al. (2007), Duffie et al. (2007), Duffie et al. (2009), and Lando and Nielson (2010). This strand of research is mainly centered around the definition and identification of a set of significant covariates influencing the timing of defaults. For our study, the clustering of default inevitably brings the challenge of using the appropriate test statistic. This problem, which is further aggravated by the sparseness of default data, led the Basel Committee on Banking Supervision (2005) to conclude that “statistical tests alone will be insufficient to adequately validate an internal rating system.” However, Blöchlinger and Leippold (2011) dispel these concerns by developing a test statistic for event forecasting, which overcomes sparseness and dependency. Their test statistic was subsequently extended to a multiperiod setting in Blöchlinger (2011). He shows that well known discrimination figures like the Gini index or the area under the ROC curve can differ in different states of the economy. When comparing the discriminatory power of two models, however, the dependence on the state of economy can be filtered out. By applying the summary statistics developed in Blöchlinger (2011) for multiple time periods, we can perform a thorough statistical analysis of the different models over time.

Not only the agencies' role in rating structured products and corporate debt has been heavily criticized, but also the role of regulators. Already before the recent crisis, White (2007) suggests

¹¹A possible solution to the inherent principal-agent problem for credit ratings is given, e.g., in Stolper (2009).

¹²See, e.g., Coval et al. (2009), Benmelech and Dlugosz (2009), Bolton et al. (2012), Gordy and Willemann (2012), Mathis et al. (2009), and Skreta and Veldkamp (2009).

a removal of the entire NRSRO structure. He furthermore levels criticism at the Securities and Exchange Commission (SEC) for its opaque policy in establishing criteria for designating NRSROs. Later, White (2010) warns against setting up new regulations to tame credit rating agencies, as it would be unclear to what extent such regulation will succeed in pressing rating agencies to make better judgments. In his conclusion, White (2010) suggests that those involved in the public debate on rating agencies should ask the question, whether financial institutions can “instead be trusted to seek their own sources of information about the creditworthiness of bonds, so long as financial regulators oversee the safety of those bond portfolios.” We believe that our contribution provides valuable guidance in answering this question.

3 Methodology

To forecast default, we consider two accounting-based models, the five components of the Z-score model of Altman (1968) and the seven components of the Zeta-score model of Altman et al. (1977), and we also consider a market-based model, the Merton (1974) distance-to-default model. In addition, we also introduce a model that takes advantage of the virtues of both these approaches, i.e., we combine accounting data with Merton’s market-based model.

3.1 The Merton Distance-to-Default Model

The Merton (1974) distance-to-default model generates a default probability for borrowers with market-traded equity. The Merton (1974) model stipulates that the equity value of a firm satisfies:

$$E_i = V_i \Phi(d_i) - e^{-r} F_i \Phi(d_i - \sigma_i), \quad (1)$$

where E_i is the market value of the i th firm’s equity, V_i is the market value of the firm’s total asset, F_i is the face amount of the firm’s debt, r is the continuously compounded risk-free interest rate, $\Phi(\cdot)$ is the cumulative Gaussian distribution function, σ_i is the volatility of the firm’s total asset, and d_i is given by:

$$d_i = \frac{\log V_i - \log F_i + r + \frac{1}{2}\sigma_i^2}{\sigma_i}.$$

Under Merton (1974)’s assumptions, the equity value is a function of the asset value of the firm. By Ito’s lemma we can show that asset volatility and equity volatility σ_i^E are related by:

$$\sigma_i^E = \frac{V_i}{E_i} \Phi(d_i) \sigma_i. \quad (2)$$

In implementing the distance-to-default model, we follow Bharath and Shumway (2008). The procedure can be described in five steps. In the first step, we estimate σ_i^E from historical stock returns data. We use historical daily returns data over a rolling window of one year to estimate σ_i^E .

In the second step, we choose a measure of the face value of the firm's debt F_i . We use the current liabilities plus one-half of long-term debt. In the third step, we collect the equity values and risk-free rates in the corresponding home country of the firm. The equity value E_i is calculated as the product of share price and the number of shares outstanding. The one-year risk free rate r approximated by the generic on-the-run government bill, note, or bond indices.¹³

While the first three steps give us values for each of the variables in equations (1) and (2), we still need the total value of the firm V_i and the volatility of firm value σ_i . In the fourth step, we compute V_i and σ_i by simultaneously solving equations (1) and (2). However, the market leverage moves up and down far too much for equation (2) to provide reasonable estimates. To solve this problem, we follow Bharath and Shumway (2008) by implementing an iterative procedure. First, we choose a starting value $\sigma_i = \sigma_i^E \frac{E_i}{E_i + F_i}$ and we use this value and equation (1) to infer the market value of each corporation's assets every day of the previous year. We then compute the implied asset return for each day and use that time series to update our estimates for σ_i and the mean return μ_i . We iterate on σ_i in this way until convergence.¹⁴ For a small number of companies and months, the iterative procedure does not converge within 14 iterations. As in Bharath and Shumway (2004), we drop the nonconverging firm-months. For every company i for which the numerical solution converges, we compute the distance-to-default DtD_i as:

$$DtD_i = \frac{\log V_i - \log F_i + \mu_i - \frac{1}{2}\sigma_i^2}{\sigma_i}. \quad (3)$$

Finally, in the fifth step we transform the (negative) distance-to-default into a default probability PD_i for company i by the link function $G(\cdot)$ from \mathbb{R} onto $(0, 1)$:

$$PD_i = G(-DtD_i) = G\left(-\frac{\log V_i - \log F_i + \mu_i - \frac{1}{2}\sigma_i^2}{\sigma_i}\right). \quad (4)$$

If the assumptions of Merton (1974) hold true, i.e., if log returns are Gaussian and asset values reflect all public available information in a frictionless market, then the link function $G(\cdot)$ is necessarily given by the standard Gaussian distribution function $\Phi(\cdot)$. In this case, the probability p_i is calibrated and the most powerful default predictor among all $\sigma(V_i)$ -measurable functions. Indeed, Bharath and Shumway (2008) use the Gaussian distribution to convert distances-to-default into default probabilities. For our implementation of the Merton model, we use an empirical estimate of the link function $G(\cdot)$.

¹³For each of the following currency we have interest rates available: AUD, BRL, CAD, CHF, EUR, GBP, HKD, IDR, INR, JPY, KRW, MXN, MYR, NOK, NZD, SEK, SGD, THB, USD. For the remaining currencies we use the interest rates in brackets: ARS (USD), CLP (USD), CZK (EUR), DKK (EUR), INR (USD), KWD (USD), TRY (EUR), TWD (JPY).

¹⁴We iterate until the absolute difference in adjacent σ_i 's is less than 0.001.

3.2 Augmenting the credit-scoring models

In the appendix, we briefly review the Z-score of Altman (1968) and the Zeta-model of Altman et al. (1977) and we provide a closer description of the variables used in our calculations. The Z-score uses five balance-sheet ratios: The ratio of working capital to total assets Z_1 , the ratio of retained earnings to total assets Z_2 , the ratio of earnings before interest and taxes to total assets Z_3 , the ratio of market value of equity/book value of total debt Z_4 , and the ratio of sales to total assets Z_5 . The Zeta-score model contains seven variables: Earnings before interest and taxes to total assets ζ_1 , stability of earnings ζ_2 , debt service ζ_3 , cumulative profitability ζ_4 , liquidity ζ_5 , capitalization ζ_6 , and company size ζ_7 .

To improve the predictive power of our replicate of Bharath and Shumway's (2004) Merton model, we augment the model by the seven components of the Zeta-model as predictor variables. As our empirical tests show, the Z-score model of Altman (1968) is dominated by the Zeta-model (see Table 4). Therefore, we only consider the latter model for our extension and we leave aside the Z-model altogether.

For the augmented model, we use a Gaussian link function $\Phi(\cdot)$. In addition, we apply rank and power transformations to map the distance-to-default into a default probability.¹⁵ In particular, the generic version of the augmented model has the form

$$PD_i = G(DtD_i) = \Phi(\alpha + \beta R(DtD_i) + \gamma R(DtD_i)^p + \delta \sum_{j=1}^7 \eta_j R(\zeta_{ij})^q), \quad (5)$$

where $R(DtD_i)$ is the rank transformed distance-to-default variable. The power parameters p and q will be fixed, whereas α , β , γ , δ , and η_j , $j = 1, \dots, 7$, are parameters to be estimated. The variables ζ_{ij} are the Zeta variables of company i . When we adjust our default probability estimates for the business cycle, we make the parameter α time dependent, i.e., we replace in equation (5) the variable α by some α_t .

4 Data Set

Our data set includes all S&P-rated non-financial corporations between 1982 and October 2011, for which we collect equity and debt data from Bloomberg. All empirical results presented refer to ratings by S&P. However, we are not aware of any reason why the conclusions we draw from our empirical analysis for S&P ratings should not apply for ratings by Moody's and Fitch. Therefore, our discussions and findings can be generalized to the ratings of Moody's and Fitch.

¹⁵Ranking transformations are often used to transform data that do not meet the assumptions of normality. E.g., one method replaces each original data value by its rank. This procedure has been recommended as being robust to non-normal errors, resistant to outliers, and highly efficient for many distributions. See, e.g., Conover and Iman (1981).

A firm is default if it is downgraded to D by S&P within a year. At the end of each month, we observe for each firm the predictor variables and the S&P rating. Our objective is to predict whether the firm defaults within one year. The distance-to-default is a market-based measure and changes every day. In contrast, most of the variables of the Z-score and Zeta-score are based on balance sheet data, which are available at most quarterly. For balance sheet data, we use the latest available report. We do not assume that the reporting data is available immediately at the reporting date, but with a time lag. We therefore restrict the data set on the data available in October for every year and firm. Hence, our analysis is truly out-of-sample. We summarize the data set used for our empirical study in Table 1.

[TABLE 1 about here]

We divide our data into an estimation sample and a validation sample. The estimation sample ranges from 1982 to 1999 and serves to estimate the parameters of the default forecasting models. The validation sample ranges from 2000 to 2011 and allows us to test the performance of the different default forecasting models and to benchmark them against the agencies' credit ratings. Our S&P rating data ends in October 2011, hence the last date for which we can analyze whether a given company defaults within one year is October 2010. To obtain correct statistics, we look at a single month for every company and year, i.e., we include the S&P rating, accounting data, and distance-to-default by the end of October for every company and year. In total, the estimation sample contains 7,684 firm-months, the validation sample contains 16,727 firm-months, for which we are able to compute the distance-to-default.

Moody's KMV entertains some version of the Merton (1974) distance-to-default model. Fink (2003) contains a table with Moody's default probabilities and volatilities for 95 companies as of August 2000. As a sanity check, Bharath and Shumway (2008) compare estimated volatilities and default probabilities of 80 companies with these numbers. The rank correlation between their estimates and Moody's is 80% for the default probabilities and 57% for the volatilities, respectively.

We perform the same sanity check with our data. In total, we have 78 companies for which we can compute the rank correlation between our estimates and Moody's. For the default probabilities, we get a rank correlation of 83.4% and 90.3% for the volatility estimates, respectively. Table 2 summarizes these results. Hence, we can safely conclude that our method in implementing a distance-to-default model is similar to the version of Bharath and Shumway (2008).

[TABLE 2 about here]

5 Statistical Analysis

We split our statistical analysis into the analysis of the estimation sample and the validation sample. The estimation sample from 1982 to 1999 is used for the in-sample estimation. The validation sample period from 2000 to 2010 is then used for the out-of-sample analysis.

5.1 In-Sample Estimation

For our analysis, we specify different calibration functions. As a first candidate for a specification of the calibration function, we use the original structural approach of Merton (1974) and label it by **M1**. It generates the default probability PD as¹⁶

$$\mathbf{M1} \quad : \quad \text{PD} = \Phi(-DtD),$$

where $\Phi(\cdot)$ denotes the cumulative distribution function of a standard Gaussian variable, which transforms the distance-to-default DtD into a one-year probability of default.

We also estimate a model, where the distance-to-default enters the score linearly. In particular, we estimate the alternative calibration function labeled **M2** as,

$$\mathbf{M2} \quad : \quad \text{PD} = \Phi(\alpha + \beta DtD) = \Phi(-1.57 - 0.34DtD).$$

In Table 3, the large t -values indicate that the constant α becomes significantly different from zero and that the sensitivity parameter β is significantly different from one. Thus, the calibration function **M2** outperforms **M1**.

Since it is a well-known fact that asset returns are non-Gaussian and fat-tailed, we expect that rank and power transformations applied to the distance-to-default variable DtD will improve the goodness-of-fit. Therefore, we estimate a calibration function of the form

$$\text{PD} = \Phi(\alpha + \beta R(DtD) + \gamma R(DtD)^p),$$

for different and fixed values of p , where $R(DtD)$ denotes the rank-transformed distance-to-default variable. Indeed, as we see from Table 3, the rank-transformed distance-to-default variable with $p = 30$, i.e.,

$$\mathbf{M3} \quad : \quad \text{PD} = \Phi(-4.49 + 2.64R(DtD) + 1.63R(DtD)^{30}),$$

has the highest likelihood value of all Merton-based model specifications. Hence, an increase in the distance-to-default variable increases the default probability more when the distance-to-default is already high. As we report in Table 3, the transformation of $R(DtD)$ to the power of 30 is more powerful than the transformations to the power of 10 (model **M4**) and 50 (model **M5**).¹⁷

¹⁶For notational convenience, we drop the index i from the subsequent formulas.

¹⁷We have also tested the model with other values for p , but we do not report the results here.

[TABLE 3 about here]

In Table 4, we report the results from the estimation of probit models for the Z-score and the Zeta-score variables. As we have already mentioned, the Zeta-score model statistically outperforms the older Z-score model. We also note that with rank transformed variables, we obtain a better fit than by simply winsorizing the variables. Some Zeta variables are statistically insignificant. Therefore, we exclude these variables from the subsequent analysis and we only use debt service (ζ_3), cumulative profitability (ζ_4), capitalization (ζ_6), and company size (ζ_7).

[TABLE 4 about here]

To limit the influence of outliers, we apply a rank transformation to the original financial ratios. Furthermore, we find that an additional square-root transformation allows for a better fit of the data. Figure 1 illustrates the transformations of the four Zeta variables used in our model. As an example, in Panel B we plot the transformation for the Zeta variable ‘cumulative probability’. If a corporation has no retained earnings then the twice transformed cumulative profitability variable is 48.3%, compared to the rank transformed cumulative profitability variable with value 22.4%. These values are then used for the assessment of the default probability. In particular, we observe that the following specification of the Zeta-score model **Z1** provides the best fit for the estimation period:

$$\begin{aligned} \mathbf{Z1} \quad : \quad \text{PD} &= \Phi \left(-0.02 - 0.61\sqrt{R(\zeta_3)} - 0.95\sqrt{R(\zeta_4)} - 1.84\sqrt{R(\zeta_6)} - 0.93\sqrt{R(\zeta_7)} \right) \\ &= \Phi(\zeta), \end{aligned}$$

where ζ is the aggregated Zeta-score and $R(\zeta_3)$, $R(\zeta_4)$, $R(\zeta_6)$, and $R(\zeta_7)$ are the rank transformed debt service ratio, cumulative profitability, market capitalization, and company size, respectively. We find that an additional power transformation to the aggregated Zeta-score ζ does not significantly change the log-likelihood value. Thus, the aggregated Zeta-score enters our model without additional transformation. The comparison of the likelihood ratios in Table 3 shows a significant statistical superiority of our model compared to both the Altman’s Zeta-score and Merton’s distance-to-default, while Merton’s distance-to-default models outperforms Altman’s Zeta-score.

As it has already been recognized by Shumway (2001), the combination of accounting-based and market-based data might enhance the predictive power of a company’s default event, if both the accounting variables and the market-based data convey company-specific credit risk information that is not subsumed by each other. From Figure 2, we observe that the Zeta-score and the distance-to-default are positively correlated, but not perfectly so. Spearman’s rank correlation shows a coefficient of 61.3% in the estimation sample (Panel A) and 53.9% in the validation sample (Panel B). As expected, the majority of defaulters is concentrated in the tenth deciles both in the estimation sample and out of the estimation sample. Hence, as correlation is not perfect, there is hope for improvement when both variables are combined into an augmented default risk measure. Therefore, we implement different combinations of market-based and accounting variables, all nested as special cases of the generic model in equation (5). We use the following model specification, which we call

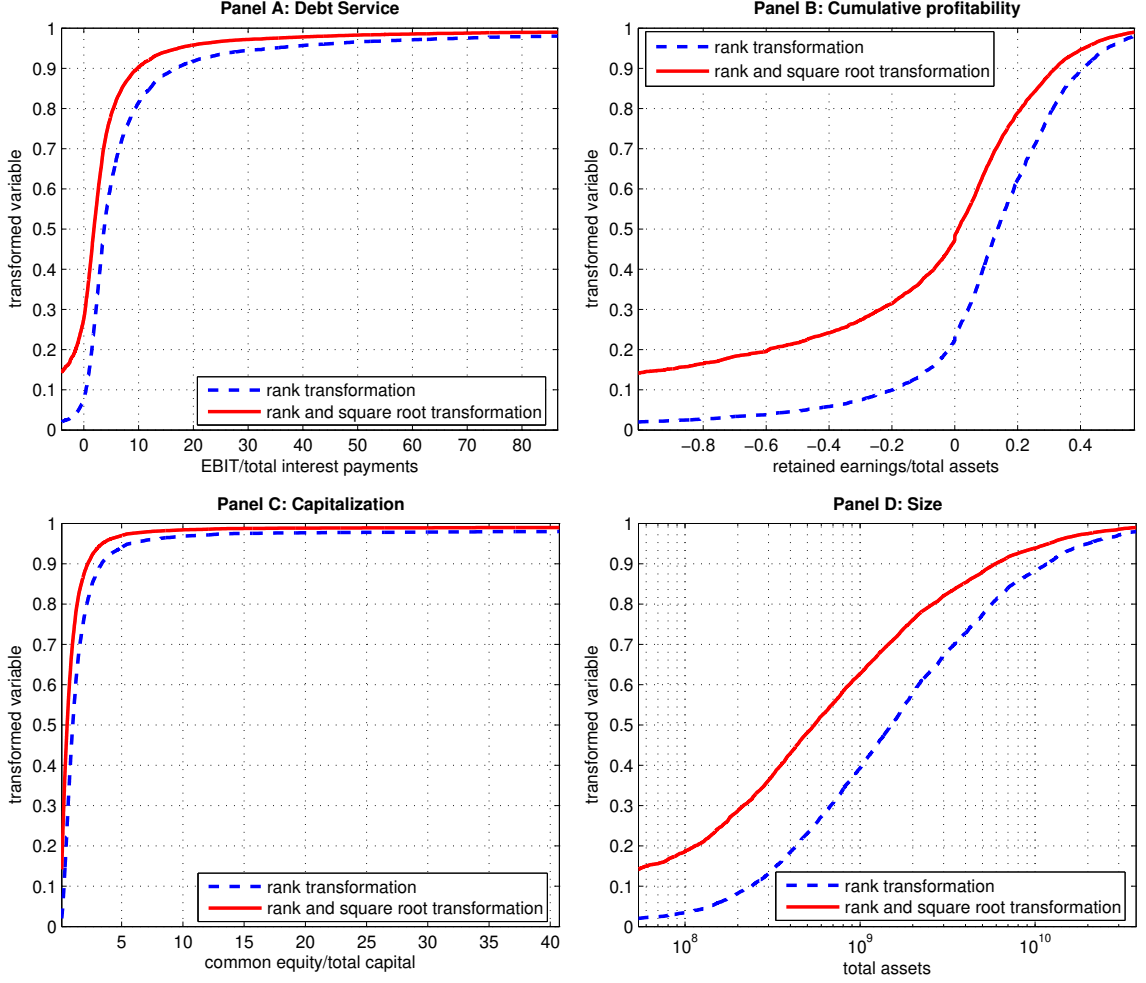


Figure 1: RANK AND SQUARE ROOT TRANSFORMATION

To limit the influence of outliers, we apply a rank transformation to the original financial ratios. To obtain a better fit of the data, we further apply a square root transformation to the rank transform of the four Zeta variables used in our model. The four transformation functions are estimated on the estimation sample and then applied on the validation sample.

the **MZ1** calibration function:

$$\mathbf{MZ1} : \quad \text{PD} = \Phi \left(-3.61 + 2.20R(DtD) + 1.40R(DtD)^{30} + 0.22\zeta \right). \quad (6)$$

With the specification in equation (6), we obtain a likelihood value significantly higher than the stand-alone distance-to-default model and Zeta-score, respectively. Furthermore, as the t -value for the Zeta-score suggest, the information content of the Zeta-score improves the overall goodness-of-fit.

In addition, we also use a business-cycle adjusted default probability, which we define as

$$\mathbf{MZ2} : \quad \text{PD} = \Phi \left(\alpha_t - 3.61 + 2.20R(DtD) + 1.40R(DtD)^{30} + 0.22\zeta \right). \quad (7)$$

The coefficient α_t is chosen at each point in time t so that the population has the same default

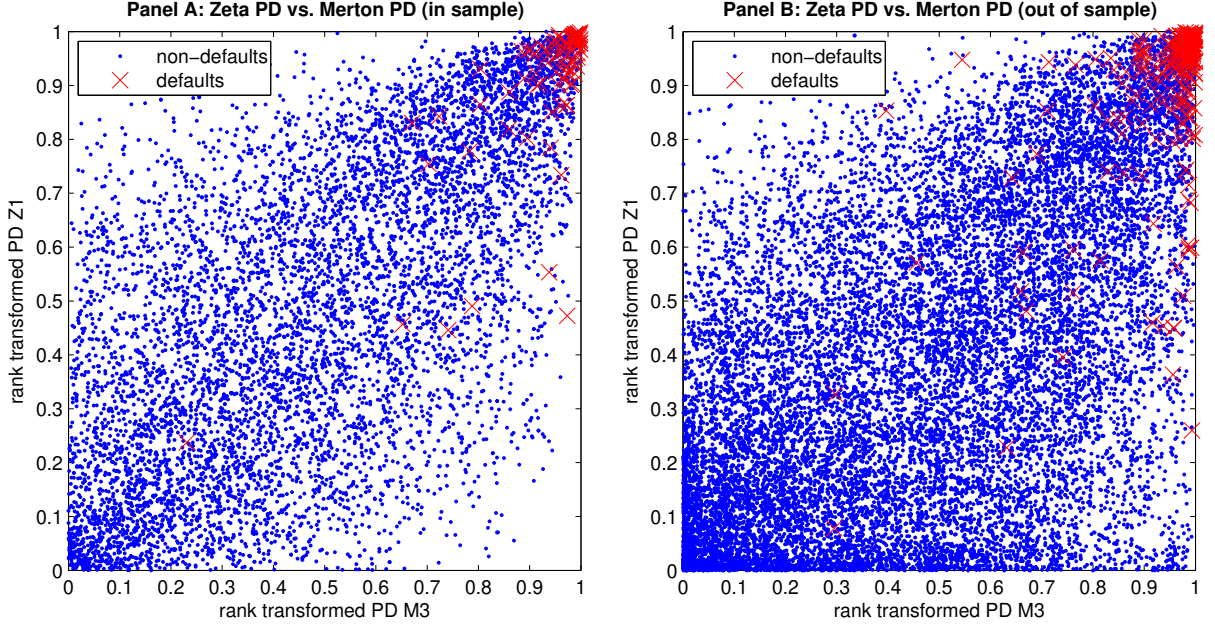


Figure 2: JOINT DISTRIBUTION ZETA-SCORE AND DISTANCE-TO-DEFAULT

The Zeta-score and the distance-to-default are correlated. Spearman’s rank correlation shows a coefficient of 67.6% in the estimation sample and 57.8% in the validation sample. We plot the estimation sample in Panel A and the validation sample in Panel B.

probability. Hence, the default probability in equation (7) corrects the default probability obtained in equation (6) for potential business-cycle effects. By filtering out these effects, we obtain a through-the-cycle forecast of default probabilities. By construction, the mean of the adjusted default probability is constant over time. In contrast, the default probability as defined in (6) would become higher in a recession and lower in a boom.

Finally, for our out-of-sample study, we require the default probability implied by the S&P rating classes. For their estimation, we use simple count statistics, i.e., we divide the number of defaulters by the number of observations per class during the estimation sample. To have strictly positive values for the default probabilities and to have a strictly monotone calibration function, we assign an estimate by interpolation to the highest rating classes.

[TABLE 5 about here]

For clarification, we illustrate the construction of the default probabilities in Table 5. These are calculated for the first quarter of each year. The first column in Table 5 lists the different companies. The variable ‘Def’ indicates whether the company is in default state, which is the case when ‘Def’ equals one. In the subsequent columns, we list the values of the four Zeta variables and the value of the aggregated Zeta variable (ζ), followed by the distance-to-default variable (DtD) and its rank transform. We include the S&P rating and list in the following column the default probabilities

implied by the agency’s rating. We then list the resulting default probabilities from our Zeta-score model **Z1**, the rank and power transformed Merton model **M3**, the hybrid point-in-time model **MZ1**, and the hybrid through-the-cycle forecast **MZ2**.

To further illustrate the construction of default probabilities, we plot the calibration functions we use for our out-of-sample validation in Figure 3. Basically, we have to map the rating scores from the S&P rating, the distance-to-default, and the Zeta-score into a default probability. For the S&P-rating we obtain this mapping by using simple count statistics, see Panel C in Figure 3. For the models **Z1**, **M3**, and **MZ1**, we apply power and rank transformations to the scores before calibrating the corresponding function to default data. The model **Z1** uses a rank and power transformed Zeta-score as input. Model **M3** uses two variables as input, a rank and power transformed distance-to-default and a rank transformed distance-to-default, respectively. We plot the transformed scores and the resulting default probabilities as a function of the score, for model **Z1** in Panel A and in Panel B for model **M3**. Model **MZ1** takes transformations of both the Zeta-score and the distance-to-default. Panel D shows the default probabilities of model **MZ1** as a function of the distance-to-default, given that the included Zeta-score is excellent (first percentile), moderate (the median Zeta-score), or disastrous (the Zeta-score is at the 99th percentile).

5.2 Out-of-Sample Validation

Before we discuss the aggregated results to evaluate the forecasting performance of the different models, we take a look at some specific companies to get some intuition about the model-implied default probabilities.

Figure 4 shows the time-series of one-year default forecasts for four defaulted companies, namely of Enron, Parmalat, Delphi, and General Motors. The performance of S&P is mixed at best. They were usually slow in down-grading a subsequent defaulter. Our observation is consistent with several surveys which show that market participants believe that rating agencies are slow in responding to changes in credit quality.¹⁸ The agencies usually reply to this critique that their ratings are “through-the-cycle” with the intention of measuring credit quality over long investment horizons. Rating stability, however, is often in conflict with rating timeliness. In the end, the default prediction performance should guide us for finding the optimal balance between these two criteria.

From Figure 4, we observe that, in contrast to the S&P credit ratings, our default probability anticipated much faster the imminent defaults of Enron, Parmalat, Delphi, and General Motors. For Enron, S&P did not change the rating until days before the corporation declared bankruptcy (Panel A). However, one of the reasons why for Enron the default probability based on the Merton model **M3** has increased may be due to a falling business cycle and a corresponding general increase of default likelihoods. Since S&P is by their own definition through-the-cycle, one may expect that it

¹⁸See, e.g., Association for Financial Professionals (2002)

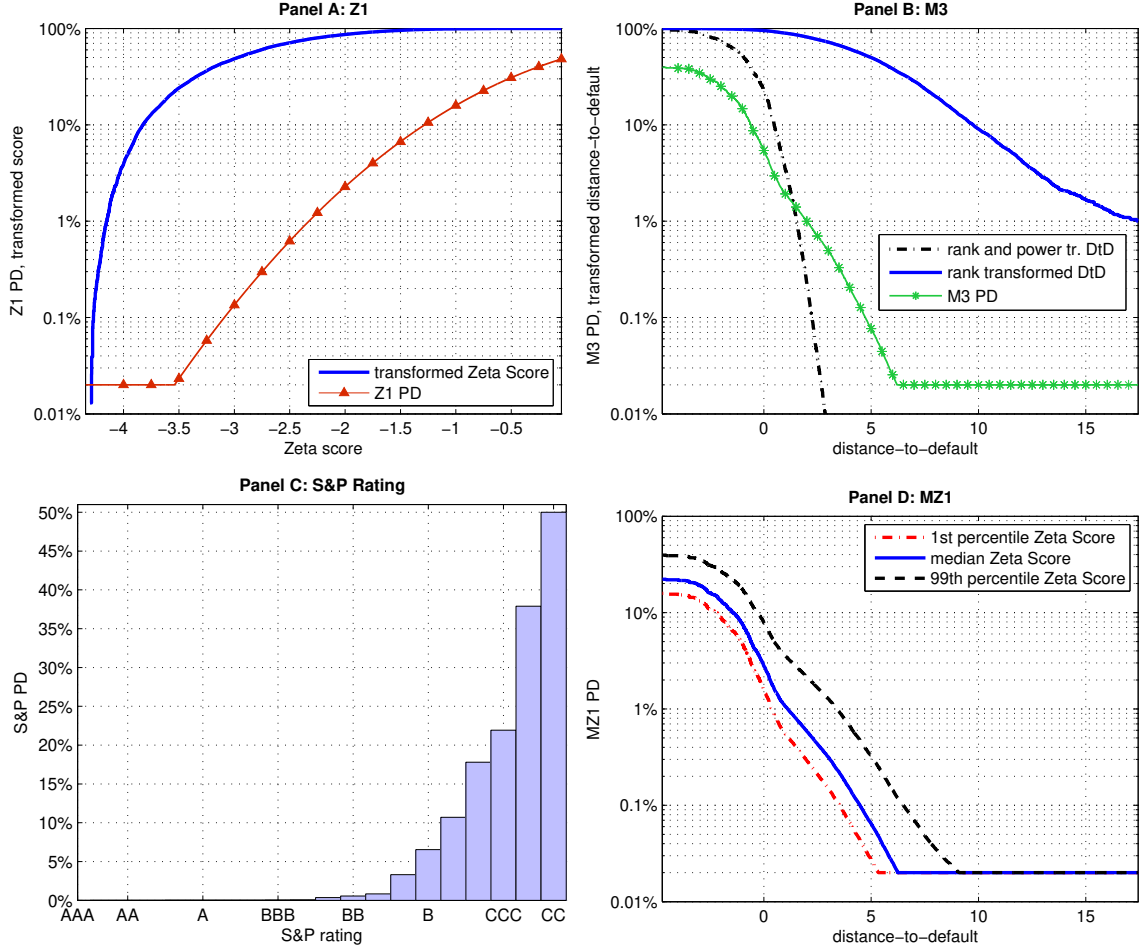


Figure 3: CALIBRATION FUNCTIONS

For the models **Z1**, **M3**, and **MZ1**, we apply transformations to the scores before calibrating the corresponding function (the mapping) to default data. The model **Z1** uses a rank and power transformed Zeta-score as input (Panel A). Model **M3** uses a rank and power transformed distance-to-default and a rank transformed distance-to-default (Panel B). Model **MZ1** takes transformations of both the Zeta-score and the distance-to-default as input (Panel D). Panel C shows the implied default probabilities of S&P credit ratings obtained by simple count statistics.

does not anticipate the economic downturn, but it must anticipate the relative worsening of Enron's individual credit standing in comparison to other companies. However, even our through-the-cycle model **MZ2** picks up the worsening of Enron's creditworthiness already at the end of 2000. For Parmalat, the story is similar (Panel B). Parmalat defaulted in December 2003 and was the biggest European default at that time, but the S&P rating for Parmalat at the end of November 2003 was still unchanged. The S&P record for the defaulted car parts supplier Delphi is a bit more favorable (Panel C). However, the Zeta-score **Z1** shows the best performance there. The Zeta-score **Z1** correctly generated low probabilities of default before the end of 2004 and then increased the default risk significantly during the last year of Delphi's life. Compared to the Merton model **M3**,

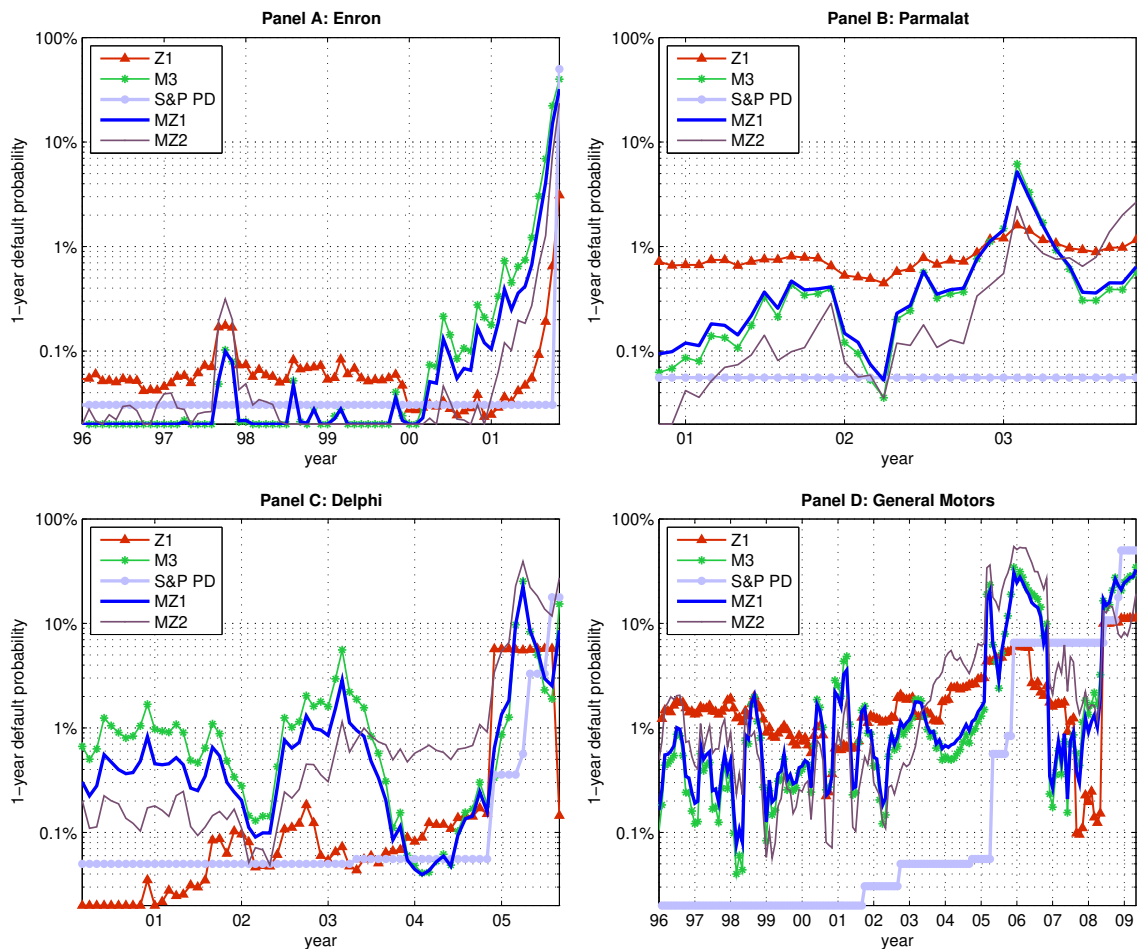


Figure 4: DEFAULTED COMPANIES

The figure shows the one-year-ahead default probability forecasts for defaulted companies, namely of Enron, Parmalat, Delphi, and General Motors. The estimation sample period used for calibrating the models ranges from 1982 to 1999. The forecasts starting from 2000 are all out-of-sample estimation.

the Zeta-score **Z1** did not produce false default signals during the performing years of Delphi. For General Motors (Panel D), the S&P ratings have correctly anticipated financial distress for the company. This is reflected by the fact that they have gradually decreased their rating until General Motors' default. However, also our alternative models seem to perform well in the case of General Motors.

While the focus in Figure 4 is on companies that actually defaulted, we may now want to switch focus on companies that have not defaulted. In Figure 5, we plot the default probability forecasts for Fiat, General Electric, Coca-Cola, and Mc Donald's. We observe that the Merton model **M3** may produce more false alarms than the other forecasting approaches. Especially for the three companies General Electric, Coca-Cola and Mc Donald's, S&P ratings and our adjusted default probability based on **MZ2** do not react to changes in the business cycle as highlighted in Figure 5.

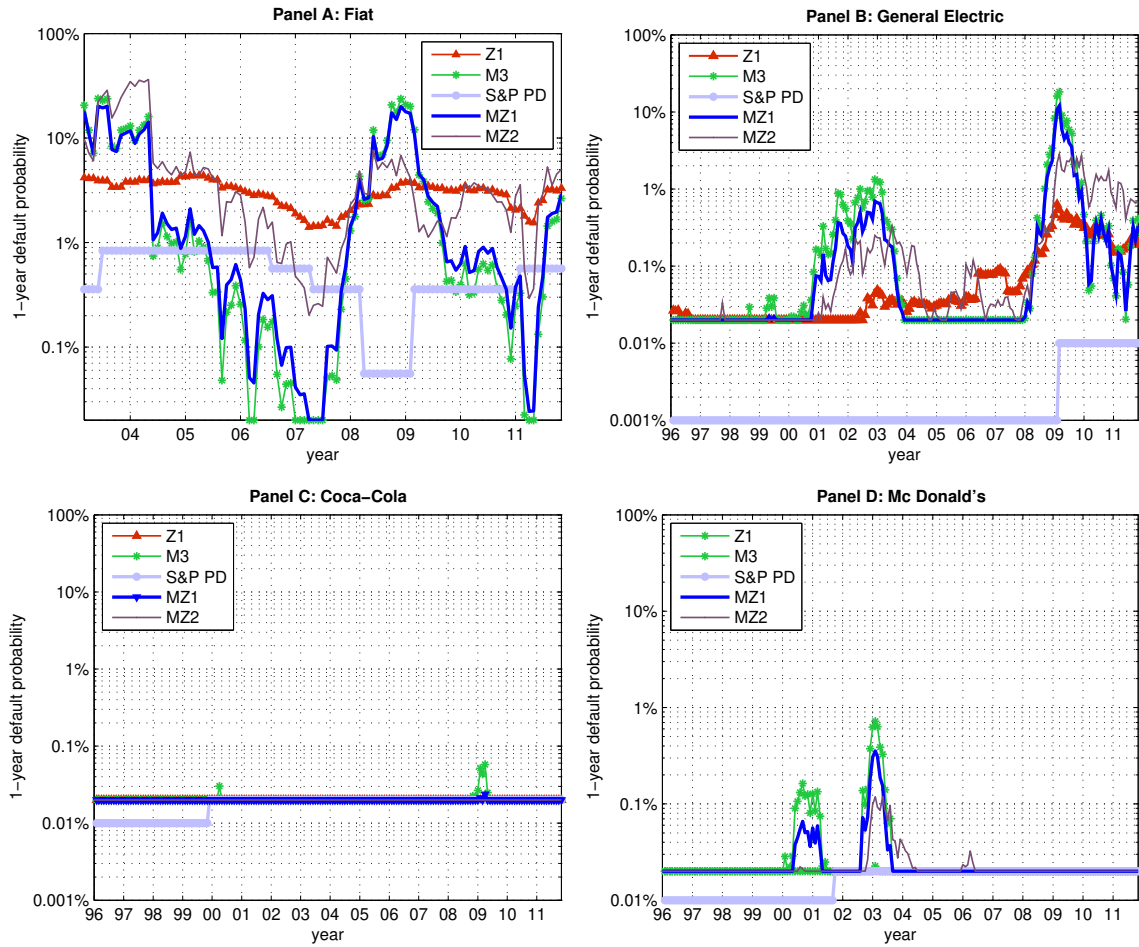


Figure 5: NON-DEFAULTED COMPANIES

The figure shows the one-year-ahead default probability forecasts for non-defaulted companies, namely of Fiat, General Electric, Coca-Cola, and Mc Donald's. The estimation sample period used for calibrating the models ranges from 1982 to 1999. The forecasts starting from 2000 are all out-of-sample estimation.

In recessionary periods, the probability of default based on models **M3**, **Z1**, and **MZ1** is higher on average as during expansionary periods of the business cycle. Merton's model **M3** produces stronger false-alarm signals in the case of General Electric (Panel B) and Mc Donald's (Panel D), relative to our default forecasts **MZ2**. The default probability based on **M3** created some spikes even though the two companies never defaulted on their debt. We observe also two spikes in the time-series of Merton's default probability **M3** for Coca-Cola (Panel C). The S&P rating of Coca-Cola, however, remained constant. On the other hand, S&P was rather slow in adjusting the rating of Fiat (Panel A). With Fiat the track record of S&P is rather dismal. In 2007 the default probability based on **MZ2** for Fiat was at a historical low. Merton's **M3** default probability implied almost no default risk. S&P followed with some delay in 2008, but it is during that time that the credit outlook of Fiat deteriorated significantly due to the financial market crisis and the anticipated recessionary period.

As a consequence, S&P reversed its erstwhile credit upgrade at the beginning of 2009, i.e., exactly at a time when the credit outlook of Fiat improved again.

To test the overall performance of the various forecasting approaches, we have to look at all companies in the validation sample simultaneously. We may have a first glimpse at the performance of the different models when we look at Table 6. We can assess the discriminatory power of the different models as follows. We order the companies according to their default probabilities and analyze the number of defaults included in a given fraction of the population. For instance, we see that among the first nine decile of S&P ratings, still 25.54% of all defaults are included. While for the augmented model **MZ1**, only 13.39% are included. Hence, approximately seven eighths of all defaulters are found in the decile with the worst credit ratings. Even if, for fairness, we look at the through-the-cycle version of our model (**MZ2**), we find that 16.54% of all defaulters are included in the first nine deciles. Therefore, we can conclude that the difference between the S&P ratings and **MZ1** indicates a superior discriminatory power of the latter model. The superiority even holds under the adjustment of business cycle effects.

[TABLE 6 about here]

We can collect the information from the above discriminatory analysis using the Lorenz curve. The Lorenz curve is a graphical tool to illustrate the discrimination ability of probability forecasts. We plot in Figure 6 the Lorenz curves of the different models. The area above the Lorenz curve serves as summary statistic of the models' discriminatory power. For the validation period, our probability forecasts with the **MZ1** model generate an area above the Lorenz curve of 94.50%, compared with an area of 92.49% for Altman's Zeta-score **Z1**, 93.77% for Merton's default probability **M3**, and 91.84% for the S&P credit ratings. Even when we adjust our forecasts **Z2** with respect to the business cycle, we still have a higher area above the Lorenz curve than we would obtain with S&P's credit ratings, i.e., 93.09% versus 91.7%. Thus, the out-performance of our simple credit scoring model **MZ1** is not purely based on the fact that the default probabilities increase (decrease) on average during recessionary (expansionary) periods.

However, to make a precise statement about the potential out-performance of the model **MZ1** with respect to the other models, we require a sound statistical test. In particular, we have two major validation aspects that we need to assess, i.e., discrimination and calibration. Models that distinguish well between borrowers who default and those who survive are said to have good discrimination. In contrast, calibration refers to the model's ability to match predicted and observed default rates. A model where the number of observed defaults align well with the model-implied number of defaults demonstrates good calibration. Blöchlinger and Leippold (2011) have recently developed a one-period test statistic under dependent observations which proved to be more powerful than existing tests, particularly when data is sparse and clustered. These features are typical for default data. The test statistic in Blöchlinger and Leippold (2011) has been extended by Blöchlinger (2011) to

a multiperiod setting test statistic for calibration and discrimination power. Defaults are typically dependent, therefore the test statistics require assumptions on the default correlations. The tests assume that default correlations are driven by a factor. Conditional on this factor, defaults are independent and Bernoulli distributed. The factor is assumed to have a Beta distribution. Now, dependency is completely specified with two parameters, the factor loading ω , and the factor volatility σ . We choose $\omega = 0.8$, and $\sigma = 0.7889$. For details regarding the choice of these parameters, and technical details of the tests, we refer to Blöchliger (2011).

[TABLE 7 about here]

Table 7 and Table 8 report the discriminatory power and the calibration tests. For the discriminatory power, we report in Table 7 the area above the Lorenz curve for the probability forecasts based on **MZ1**, the Altman’s Zeta-score **Z1**, Merton’s adjusted distance-to-default model **M3**, S&P ratings, and our business cycle adjusted model **MZ2**. In addition, we report the t -values for the hypothesis that the areas above the curve are different. We also include what we call the naive forecast, which assumes that all companies share the same probability of default. Hence, the naive forecast implies no discriminatory power at all. As expected, we see that all model-based approaches differ significantly from the naive forecast. We also observe that the forecast of our model **MZ1** is significantly superior to all other forecasts at the 1% significance level. Hence, the null hypothesis of the same discriminatory power must be rejected. For the business cycle adjusted version of **MZ1**, however, we cannot statistically prove its superiority compared to the S&P ratings at a reasonable confidence level. Therefore, our intermediate conclusion is that our through-the-cycle rating **MZ2** is at least as good as the S&P ratings, with the advantage that they are based on publicly available data only and the model is fully transparent and replicable.

So far, we have only looked at the discriminatory power of the default forecasts and we left aside the question of whether the forecasts are calibrated. Calibration means that the forecasts are unbiased. In Table 8 we report the relevant test statistics, i.e., the level, the shape, and the combined statistic. The level statistic compares the realized default rate with the predicted rate. The shape statistic depends on the difference between the realized and the expected area above the Lorenz curve. Both normalized statistics are χ^2 -distributed with one degree of freedom. The combined statistic is a summary statistic on level and shape and follows a χ^2 -distribution with two degrees of freedom. We see in Panel A of Table 8 that in each year our model **MZ1** provides calibrated forecast. S&P ratings in Panel B of Table 8, however, produce non-calibrated and hence biased forecasts in three out of eleven years (see combined p -value). We note that even with the through-the-cycle model **MZ2** we are able to generate p -values for the combined test statistic, which are all above the 5% level.¹⁹ Hence, even our model **MZ2** is calibrated. Table 8 also reports a multi-period test in the last line of Panels A, and B respectively. The multi-period statistic is the sum of the corresponding

¹⁹These additional results can be obtained by the authors.

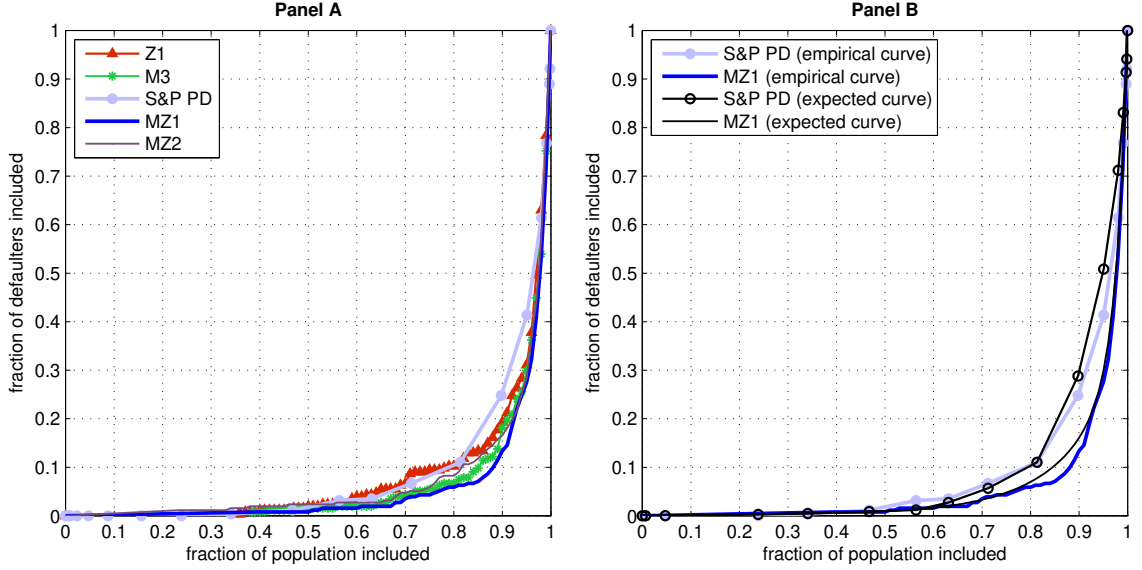


Figure 6: DISCRIMINATORY POWER AND SHAPE CALIBRATION

Our model shows the best discrimination ability to separate defaulting from non-defaulting corporations on the validation sample from 10/2000 to 10/2010. Our approach, **MZ1**, is significantly more powerful than Altman’s Zeta-score, **Z1**, Merton’s distance-to-default, **M3**, and S&P’s corporate issuer rating. Both S&P ratings and our model demonstrate good shape calibration. The shape calibration hypotheses cannot be rejected. The empirical Lorenz curves do not deviate significantly from the Lorenz curves that are expected if the null hypotheses of shape calibrated forecasts are true.

independent single period statistics and thus is χ^2 -distributed too. The degree of freedom equals the number of single-period statistics for the shape statistics, and level statistics respectively, and twice this number for the combined multi-period statistic. The null hypothesis of a level and shape calibrated rating methodology must be rejected for S&P ratings with a p-value of 0.06%, but clearly not for our default probabilities obtained with **MZ1**.

[TABLE 8 about here]

6 Conclusion

Financial legislations, some of them decades old, often require banks and corporations to rely on credit ratings in their everyday business. Given the rating agencies’ dubious role in the financial crisis, regulators and politicians are now aiming to wean the financial industry off its dependence on credit rating agencies. The Dodd-Frank act mandates that all financial agencies scan their regulations for references to credit ratings, and then remove and replace them with an appropriate alternative. In search for such an appropriate alternative for the case of corporate credit ratings, we use a combination of quantifiable credit concepts, namely Merton’s distance-to-default model and Altman’s

Zeta variables. By following such an approach, we develop a credit risk model which is, unlike credit ratings, not a black box, but transparent, replicable, and easy to implement.

For our out-of-sample analysis, we calibrate default probability forecasts using the Standard & Poor's corporate data base covering the period from October 1982 to October 1999. We then use the calibrated models to perform out-of-sample forecasts for the subsequent period, starting October 2000 and ending October 2011. We show that a combination of publicly available accounting and market data is significantly more powerful than the stand-alone concepts and significantly more powerful than the credit ratings from S&P. In addition, compared to credit ratings we not only provide an ordinal measure of default risk but a calibrated probability of default, i.e., our probability forecasts are unbiased.

Since rating agencies claim to provide a through-the-cycle assessment, the default frequency per rating class fluctuates over the business cycle. We therefore adjust our model to provide a through-the-cycle estimate. With respect to the discriminatory power, our through-the-cycle estimates are at least as good as the estimates implied by the credit ratings. However, while we find that the credit ratings are biased, we can confirm unbiasedness for our estimates. Thus, our findings can be summarized as follows: For the case of corporate credit risk, appropriate alternatives to credit ratings do exist!

References

- Altman, E.: 1968, Financial ratios, discriminant analysis and the prediction of corporate bankruptcy, *Journal of Finance* **23**(4), 589–609.
- Altman, E., Haldeman, R. and Narayanan, P.: 1977, Zeta tm analysis a new model to identify bankruptcy risk of corporations, *Journal of Banking & Finance* **1**(1), 29–54.
- Altman, E. I. and Rijken, H. A.: 2006, A point-in-time perspective on through-the-cycle ratings, *Financial Analysts Journal* **62**(1), 54–70.
- Association for Financial Professionals: 2002, *Rating agencies survey: Accuracy, Timeliness and Regulation*, www.afponline.org.
- Basel Committee on Banking Supervision: 2005, Studies on the validation of internal rating systems, *Working Paper 14*, Bank for International Settlement.
- Becker, B. and Milbourn, T.: 2011, How did increased competition affect credit ratings?, *Journal of Financial Economics* **101**(3), 493 – 514.
- Benmelech, E. and Dlugosz, J.: 2009, The alchemy of cdo credit ratings, *Journal of Monetary Economics* **56**(5), 617–634.
- Bharath, S. and Shumway, T.: 2004, Forecasting default with the KMV-Merton model, *AFA 2006 Boston Meetings Paper*.
- Bharath, S. and Shumway, T.: 2008, Forecasting default with the Merton distance to default model, *Review of Financial Studies* **21**(3), 1339.
- Blöchlinger, A.: 2010, Arbitrage-free credit pricing using default probabilities and risk sensitivities, *Journal of Banking & Finance*.
- Blöchlinger, A.: 2011, Validation of default probabilities, *Journal of Financial and Quantitative Analysis* **forthcoming**.
- Blöchlinger, A. and Leippold, M.: 2011, A new goodness-of-fit test for event forecasting and its application to credit defaults, *Management Science* **57**(3), 471 – 486.
- Bolton, P., Freixas, X. and Shapiro, J.: 2012, The credit ratings game, *Journal of Finance* **67**(1), 85–112.
- Boot, A. W. A., Milbourn, T. T. and Schmeits, A.: 2006, Credit ratings as coordination mechanisms, *Review of Financial Studies* **19**(1), 81–118.
- Cantor, R.: 2004, An introduction to recent research on credit ratings, *Journal of Banking and Finance* **28**(11), 2565 – 2573.

- Conover, W. J. and Iman, R. L.: 1981, Rank transformations as a bridge between parametric and nonparametric Statistics, *The American Statistician* **35**(3), 124–129.
- Coval, J., Jurek, J. and Stafford, E.: 2009, The economics of structured finance, *Journal of Economic Perspectives* **23**(1), 3–25.
- Crouhy, M., Galai, D. and Mark, R.: 2001, Prototype risk rating system, *Journal of Banking and Finance* **25**(1), 47–95.
- Das, S., Duffie, D., Kapadia, N. and Saita, L.: 2007, Common failings: How corporate defaults are correlated, *Journal of Finance* **62**, 93–117.
- Duffie, D., Eckner, A., Horel, G. and Saita, L.: 2009, Frailty correlated default, *Journal of Finance* **64**, 2089–2123.
- Duffie, D., Saita, L. and Wang, K.: 2007, Multi-period corporate default prediction with stochastic covariates, *Journal of Financial Economics* **83**(3), 635–665.
- Fink, R.: 2003, Where credit is due, *CFO*.
- Gordy, M. B. and Willemann, S.: 2012, Constant proportion debt obligations: A post-mortem analysis of rating models, *Management Science* **forthcoming**.
- Hilscher, J. and Wilson, M. I.: 2012, Credit ratings and credit risk, *Working paper*, Brandeis University and University of Oxford.
- Lando, D. and Nielson, M. S.: 2010, Correlation in corporate defaults: Contagion or conditional independence?, *Journal of Financial Intermediation* **19**(3), 355–372.
- Mathis, J., McAndrews, J. and Rochet, J.-C.: 2009, Rating the raters: Are reputation concerns powerful enough to discipline rating agencies?, *Journal of Monetary Economics* **56**(5), 657–674.
- Merton, R.: 1974, On the pricing of corporate debt: The risk structure of interest rates, *Journal of finance* **29**(2), 449–470.
- Partnoy, F.: 1999, The siskel and ebert of financial markets: two thumbs down for the credit rating agencies, *Washington University Law Quarterly* **77**, 619–714.
- Shumway, T.: 2001, Forecasting bankruptcy more accurately: A simple hazard model, *The Journal of Business* **74**(1), 101–24.
- Skreta, V. and Veldkamp, L.: 2009, Ratings shopping and asset complexity: A theory of ratings inflation, *Journal of Monetary Economics* **56**(5), 678–695.
- Stolper, A.: 2009, Regulation of credit rating agencies, *Journal of Banking and Finance* **33**(7), 1266 – 1273.

Summers, L.: 1985, On economics and finance, *The Journal of Finance* **40**(3), 633–635.

US Department of Justice: 1998, Comments of the United States Department of Justice in the Matter of: File No. S7-33-97 Proposed Amendments to Rule 15c3-1 under the Securities Exchange Act of 1934, *Technical report*.

Vassalou, M. and Xing, Y.: 2004, Default risk in equity returns, *The Journal of Finance* **59**(2), 831–868.

White, L. J.: 2007, A new law for the bond rating industry, *Regulation* **30**(1), 48–52.

White, L. J.: 2010, Markets: The credit rating agencies, *The Journal of Economic Perspectives* **24**(2), 211–226.

Appendix

Variables of the Z-Score Model

- Z_1 compares liquid assets to total assets. The ratio is computed as working capital (current assets minus current liabilities) over total assets.
- Z_2 quantifies cumulative profitability over time as the ratio of retained earnings over total assets.
- Z_3 illustrates the productivity of the firm's assets, abstracting from leverage and taxes. The ratio is calculated as earnings before interest and taxes (EBIT) over total assets. We compute EBIT by summing up EBIT for the most recent four quarters. EBIT is after amortization of goodwill.
- Z_4 is a measure for leverage, and the only ratio of the Z-score which contains market prices. The ratio is defined as the market value of equity over the book value of total debt. We compute the market value of equity by multiplying the number of shares with the closing price of the share. The number of shares is the number of primary common shares of all classes outstanding, net of treasury shares.
- Z_5 quantifies the sales generating ability of the firm as the ratio of sales over total assets. We consider the trailing twelve-month sales, i.e., the sum of the quarterly sales of the most recent four quarters.

Variables of the ZETA Model

The ZETA model contains the following seven variables:

- ζ_1 corresponds to the ratio Z_3 of the older Z-score model, EBIT over total assets. However, in the original ZETA model Altman et al. (1977) include total tangible assets as the denominator. Since many companies in our sample do not disclose intangibles, we use total assets instead.
- ζ_2 quantifies the stability of earnings. Altman et al. (1977) describe this measure as a normalized measure of the standard error of estimate around a ten-year trend in ζ_1 . We compute this variable as the coefficient of determination (R-squared) with the trailing twelve-month EBIT as the dependent variable, and time as the independent variable. If EBIT follows a trend, this variable is close to one, whereas it is close to zero if the time series of EBIT cannot be well explained with a trend line. We choose a five-year trend. We compute R-squared as soon as three values for ζ_1 are available (this is usually the sixth reporting quarter).
- ζ_3 measures the ability of the firm to meet its liabilities. This variable is computed as EBIT divided by total interest payments, plus working capital divided by long term debt. Total interest payments are computed as the trailing twelve-month payments, i.e. the cumulated

interest payments of the latest four quarters. Long term debt is total liabilities minus current liabilities.

ζ_4 is Z_2 of the Z-score, cumulative profitability, computed as retained earnings divided by total assets.

ζ_5 measures liquidity by the current ratio, which is current assets divided by current liabilities.

ζ_6 is common equity divided by total capital. In the numerator of this ratio the common equity is measured by market value of the total equity, rather than book value.

ζ_7 measures size by total assets. This is the only variable in the Z-score and the ZETA model which has a unit. We measure total assets in million U.S. dollars.

Tables

Table 1: DATA SUMMARY

Industry	Firms		Obs.		Def. Obs.	
	1982-1999	2000-2010	1982-1999	2000-2010	1982-1999	2000-2010
Aerospace/Automotive/Capital Goods/Metal	258	425	1,297	2,982	9	51
Consumer/Service Sector	298	446	1,624	2,983	28	45
Energy and Natural Resources	139	245	614	1,511	8	12
Forest & Building Products/Homebuilders	105	137	553	952	9	21
Health Care/Chemicals	171	280	880	1,858	8	14
High Tech/Computers/Office Equipment	135	221	686	1,273	8	9
Leisure Time/Media	185	232	738	1,517	13	29
Real Estate	8	37	14	217	0	1
Telecommunications	78	166	264	860	4	45
Transportation	69	124	341	803	6	20
Utility	135	236	673	1,771	0	7
Total	1,581	2,549	7,684	16,727	93	254

We divide our data into an estimation sample covering the period from 1982 to 1999, and a validation sample with data from 2000 to 2010. Our S&P rating data ends in October 2011, hence the last date for which we can analyze whether a given company defaults within one year is October 2010. The estimation sample contains 1,581 firms, the validation sample 2,549 firms. There are 347 firms that defaulted (def.obs.), 93 are in the estimation sample, 254 in the validation sample. We have 24,411 firm-month observations (obs.) in total.

Table 2: DATA QUALITY CHECK

	Bharath and Shumway (2008)	Our results
Rank correlation of DtD	79%	83.4%
Rank correlation of volatility	57%	90.3%
Companies with data	80	78

The table reports the rank correlation between our estimates of the asset return volatility and the estimates published by Fink (2003). Likewise for the distance-to-default variable (DtD). We also display the corresponding rank correlations reported by Bharath and Shumway (2008). Note that we have two companies less in our sample.

Table 3: IN-SAMPLE CALIBRATION

Model:	M2 PD	M4 PD	M5 PD	M3 PD	Z1 PD	MZ1 PD
α	-1.57 (-29.54)	-3.66 (-10.73)	-4.96 (-10.99)	-4.49 (-10.79)		-3.61 (-6.37)
β	-0.34 (-13.46)					
γ with $p = 1$		1.06 (2.16)	3.33 (6.25)	2.64 (5.17)		2.20 (4.09)
γ with $p = 10$		2.04 (7.79)				
γ with $p = 30$				1.63 (7.36)		1.40 (5.66)
γ with $p = 50$			1.56 (6.80)			
zeta-score					1.00 (40.12)	0.22 (2.11)
Log likelihood value	-330.31	-321.70	-324.65	-321.32	-363.48	-319.06

The table reports the result from different probit regression models to transform the two ordinal credit risk measures distance-to-default and Zeta-score into a 1-year default probability. The estimates are derived from our estimation sample covering the S&P corporate data base for the period from 1982 to October 1999. The comparison of the likelihood ratios shows a significant statistical superiority of our model compared to both the Altman's Zeta-score and Merton's distance-to-default, while Merton's distance-to-default models outperform Altman's Zeta-score.

Table 4: IN-SAMPLE ESTIMATION OF Z-SCORE AND ZETA-SCORE

	Z-score	Z-score (rank tr.)	zeta-score	zeta-score (rank tr.)	zeta-score (rank tr.)	Zeta-score (twice tr.)
Constant	-2.00 (-21.48)	-1.02 (-8.05)	2.69 (3.22)	-0.62 (-4.12)	-0.81 (-8.09)	-0.02 (-0.14)
Z_1 : WorkCap/Assets	-0.49 (-1.76)	-0.29 (-1.68)				
Z_2, ζ_4 : Retain. Earn./Assets	-0.76 (-5.32)	-1.08 (-4.43)	-3.33 (-4.35)	-0.90 (-3.27)	-0.75 (-2.75)	-0.95 (-3.48)
Z_3, ζ_1 : EBIT/Assets	-4.72 (-6.86)	-1.30 (-5.30)	0.25 (1.81)	0.34 (2.10)		
Z_4 : Market Cap./Liab.	-0.14 (-2.98)	-1.81 (-6.52)				
Z_5 : Sales/Assets	0.16 (2.54)	0.45 (2.72)				
ζ_2 : Earn. Stability			0.00 (0.47)	-0.01 (-0.04)		
ζ_3 : Debt Service			-0.53 (-3.23)	-0.80 (-3.11)	-0.88 (-3.50)	-0.61 (-2.41)
ζ_5 : Current Ratio			-0.14 (-2.91)	-0.55 (-3.27)		
ζ_6 : Capitalization			-0.55 (-5.04)	-1.64 (-6.33)	-1.62 (-6.53)	-1.84 (-8.26)
ζ_7 : Size			-0.21 (-5.17)	-1.29 (-5.97)	-1.15 (-5.44)	-0.93 (-4.38)
Log likelihood value	-407.63	-376.85	-377.82	-362.20	-375.31	-363.48

We estimate several probit regression models. To limit the influence of outliers, we first winsorize the data at the 2.5% and 97.5% quantiles to estimate the Z-score and the Zeta-score. However, based on the likelihood value the rank transformation is significantly more suitable. When the financial ratios are additionally square root transformed, we obtain even better fitting models. Following the Akaike information criterion, we choose a reduced form of the original Zeta model with rank and square root transformed financial ratios, i.e., twice transformed financial variables according to the last column. The corresponding t -values are in brackets.

Table 5: CALCULATION OF DEFAULT PROBABILITIES

Name	Date	Def.	$R(\zeta_3)$	$R(\zeta_4)$	$R(\zeta_6)$	$R(\zeta_7)$	ζ	DtD	$R(DtD)$	S&P	S&P PD	Z1 PD	M3 PD	MZ1 PD	MZ2 PD
Parmalat	4/2001	0	0.310	0.388	0.117	0.877	-2.434	4.440	0.562	BBB-	0.056%	0.747%	0.134%	0.176%	0.070%
Parmalat	4/2002	0	0.333	0.414	0.180	0.875	-2.614	5.701	0.417	BBB-	0.056%	0.447%	0.036%	0.053%	0.036%
Parmalat	4/2003	1	0.328	0.342	0.065	0.891	-2.271	1.190	0.882	BBB-	0.056%	1.158%	1.696%	1.590%	0.861%
Delphi	4/2001	0	0.681	0.298	0.610	0.947	-3.483	1.892	0.827	BBB	0.050%	0.025%	1.079%	0.519%	0.225%
Delphi	4/2002	0	0.341	0.210	0.698	0.947	-3.302	4.369	0.570	BBB	0.050%	0.048%	0.143%	0.099%	0.068%
Delphi	4/2003	0	0.628	0.171	0.554	0.948	-3.305	0.811	0.906	BBB	0.050%	0.048%	2.210%	1.119%	0.588%
Delphi	4/2004	0	0.243	0.171	0.565	0.955	-3.038	5.364	0.459	BBB-	0.056%	0.119%	0.053%	0.052%	0.611%
Delphi	4/2005	1	0.227	0.059	0.001	0.939	-1.594	-2.028	0.990	BB	0.564%	5.542%	25.287%	22.459%	39.155%
GM	4/2004	0	0.243	0.269	0.017	1.000	-1.976	2.937	0.727	BBB	0.050%	2.410%	0.514%	0.708%	4.638%
GM	4/2005	0	0.078	0.264	0.009	1.000	-1.707	-1.828	0.988	BBB-	0.056%	4.395%	23.408%	20.246%	36.265%
GM	4/2006	0	0.025	0.222	0.009	1.000	-1.563	-1.848	0.988	B	6.538%	5.904%	23.408%	21.153%	52.858%
GM	4/2007	0	0.041	0.134	0.169	0.996	-2.126	3.363	0.682	B	6.538%	1.674%	0.360%	0.484%	3.344%
GM	4/2008	0	0.035	0.059	0.899	0.995	-3.026	1.220	0.880	B	6.538%	0.124%	1.665%	1.016%	1.183%
GM	4/2009	1	0.010	0.015	0.002	0.992	-1.212	-2.342	0.992	CC	50.000%	11.283%	28.600%	27.886%	12.974%
GE	4/2006	0	0.896	0.469	0.360	1.000	-3.375	7.906	0.204	AAA	0.001%	0.037%	0.020%	0.020%	0.118%
GE	4/2007	0	0.579	0.474	0.322	1.000	-3.141	11.061	0.061	AAA	0.001%	0.084%	0.020%	0.020%	0.031%
GE	4/2007	0	0.579	0.474	0.322	1.000	-3.141	11.061	0.061	AAA	0.001%	0.084%	0.020%	0.020%	0.031%
GE	4/2008	0	0.843	0.437	0.212	1.000	-3.076	4.957	0.504	AAA	0.001%	0.105%	0.080%	0.072%	0.087%
GE	4/2009	0	0.771	0.378	0.064	1.000	-2.629	-0.530	0.967	AA+	0.010%	0.428%	9.045%	5.932%	1.778%
Insilco	4/2001	1	0.129	0.035	0.046	0.186	-1.271	-0.837	0.976	B+	3.307%	10.179%	12.835%	13.872%	8.639%

In our estimation sample, a borrower rated BBB- has a mean yearly default rate of 0.56%, BBB has a default rate of 0.50%. We apply these averages to obtain the S&P default probabilities applied to the validation sample. The other default probabilities are calculated according to the different model specifications. **Z1** is the Zeta-score model, **M3** is Merton's distance-to-default model with rank and power transformed variables, and **MZ1** and **MZ2** are the two models combining the market-based Merton model with the accounting-based Zeta-score model. **MZ1** is a point-in-time forecast and **MZ2** is a through-the-cycle forecast, i.e., this default probability forecast is business cycle adjusted.

Table 6: DISCRIMINATORY POWER

Fraction of population included	Fraction of defaulters included				
	Z1	M3	S&P	MZ1	MZ2
25.00%	0.28%	0.51%	0.04%	0.50%	1.18%
50.00%	1.97%	1.57%	1.85%	0.79%	2.36%
75.00%	9.45%	5.51%	8.31%	4.33%	5.91%
80.00%	10.24%	6.69%	10.46%	5.91%	8.27%
85.00%	13.39%	9.84%	17.04%	7.09%	11.81%
90.00%	19.28%	18.09%	25.54%	13.39%	16.54%
95.00%	31.10%	29.84%	41.15%	27.55%	28.32%
96.00%	37.79%	36.22%	47.63%	32.28%	36.61%
97.00%	49.62%	44.89%	54.33%	40.90%	44.11%
98.00%	63.03%	53.91%	61.03%	53.20%	57.15%
99.00%	78.32%	74.55%	75.18%	71.62%	79.10%

The table reports the discriminatory power of the different credit rating models. The companies are ranked in descending order according to their credit quality. The first column reports the fraction of the population included. For a given fraction, we count the number of defaults. We list the corresponding numbers for the different models in the subsequent columns.

Table 7: DISCRIMINATORY POWER: *t*-VALUES IN CROSS TESTS

	Area above Lorenz curve	vs. naive	vs. Z1	vs. M3	vs. S&P	vs. MZ1	vs. MZ2
Z1	92.49%	56.957	–	-1.936	0.782	-3.463	-0.766
M3	93.77%	65.438	1.936	–	2.255	-6.092	1.637
S&P	91.84%	58.854	-0.782	-2.255	–	-3.257	-1.311
MZ1	94.50%	72.334	3.463	6.092	3.257	–	3.243
MZ2	93.09%	53.704	0.766	-1.637	1.311	-3.243	–

In the out-of-sample statistical analysis our default forecast **MZ1** approach obtains the highest area above the Lorenz curve, i.e., our model provides the most powerful forecasts to separate defaulters from non-defaulters. The *t*-values versus the naive model, Altman’s Zeta-score **Z1**, Merton’s distance-to-default **M3**, and versus our through-the-cycle approach **MZ2** are all greater than 2.576, i.e., highly significant at the 1% significance level. All approaches demonstrate discriminatory power, i.e., they are all highly significantly more powerful than naive forecasts. S&P ratings are not able to outperform Altman’s Zeta-score.

Table 8: CALIBRATION TESTING OF ONE-YEAR DEFAULT PROBABILITIES

<i>Panel A: MZ1 PD</i>												
time period	obs.	def.	exp. PD	real. PD	exp. area	real. area	level	shape	comb.	level <i>p</i> -value	shape <i>p</i> -value	comb. <i>p</i> -value
10/2000 – 10/2001	1,213	45	2.29%	3.71%	0.909	0.914	1.105	0.075	1.180	0.2931	0.7842	0.5542
10/2001 – 10/2002	1,310	45	2.42%	3.44%	0.916	0.926	0.676	0.418	1.095	0.4108	0.5179	0.5785
10/2002 – 10/2003	1,360	29	2.37%	2.13%	0.916	0.948	0.010	2.490	2.500	0.9219	0.1146	0.2865
10/2003 – 10/2004	1,551	14	0.40%	0.90%	0.941	0.952	2.688	0.111	2.798	0.1011	0.7395	0.2468
10/2004 – 10/2005	1,653	15	0.50%	0.91%	0.953	0.958	1.590	0.024	1.614	0.2074	0.8765	0.4462
10/2005 – 10/2006	1,666	6	0.36%	0.36%	0.957	0.990	0.081	0.484	0.566	0.7757	0.4864	0.7537
10/2006 – 10/2007	1,724	6	0.34%	0.35%	0.952	0.992	0.086	0.606	0.692	0.7694	0.4362	0.7075
10/2007 – 10/2008	1,642	15	0.49%	0.91%	0.957	0.925	1.712	1.226	2.938	0.1907	0.2682	0.2302
10/2008 – 10/2009	1,619	52	3.59%	3.21%	0.876	0.896	0.008	1.105	1.113	0.9283	0.2931	0.5731
10/2009 – 10/2010	1,518	19	0.77%	1.25%	0.886	0.946	1.106	2.820	3.926	0.2930	0.0931	0.1404
10/2010 – 10/2011	1,471	8	0.31%	0.54%	0.927	0.982	1.393	1.281	2.674	0.2379	0.2577	0.2626
10/2000 – 10/2011	16,727	254	1.20%	1.52%	0.941	0.945	1.669	0.506	2.175	0.1964	0.4769	0.3371
multi-period test	16,727	254	1.20%	1.52%	0.941	0.945	10.45	10.64	21.10	0.4900	0.4738	0.5148
<i>Panel B: S&P PD</i>												
10/2000 – 10/2001	1,213	45	1.65%	3.71%	0.913	0.912	2.731	0.007	2.738	0.0984	0.9330	0.2544
10/2001 – 10/2002	1,310	45	1.61%	3.44%	0.918	0.865	2.411	13.116	15.527	0.1205	0.0003	0.0004
10/2002 – 10/2003	1,360	29	1.71%	2.13%	0.924	0.939	0.378	0.681	1.059	0.5389	0.4092	0.5890
10/2003 – 10/2004	1,551	14	1.53%	0.90%	0.919	0.904	0.282	0.296	0.577	0.5957	0.5865	0.7492
10/2004 – 10/2005	1,653	15	1.33%	0.91%	0.912	0.923	0.101	0.148	0.250	0.7506	0.7000	0.8827
10/2005 – 10/2006	1,666	6	1.32%	0.36%	0.913	0.897	3.536	0.120	3.656	0.0601	0.7291	0.1608
10/2006 – 10/2007	1,724	6	1.44%	0.35%	0.910	0.984	4.945	2.734	7.679	0.0262	0.0982	0.0215
10/2007 – 10/2008	1,642	15	1.57%	0.91%	0.904	0.915	0.299	0.167	0.466	0.5844	0.6828	0.7921
10/2008 – 10/2009	1,619	52	1.58%	3.21%	0.906	0.931	2.129	2.844	4.973	0.1445	0.0917	0.0832
10/2009 – 10/2010	1,518	19	2.02%	1.25%	0.908	0.973	0.206	7.396	7.602	0.6499	0.0065	0.0223
10/2010 – 10/2011	1,471	8	1.64%	0.54%	0.901	0.967	2.123	3.061	5.184	0.1451	0.0802	0.0749
10/2000 – 10/2011	16,727	254	1.57%	1.52%	0.912	0.918	0.008	0.919	0.928	0.9269	0.3377	0.6289
multi-period test	16,727	254	1.57%	1.52%	0.912	0.918	19.14	30.57	49.71	0.0586	0.0013	0.0006

The validation sample contains 16,727 observations and 254 defaults from 11/2000 to 10/2011. The level statistic compares the realized default rate (real. PD) with the expected rate (exp. PD). The shape statistic depends on the difference between the realized (real. area) and the expected area (exp. area) above the Lorenz curve. The combined statistic is a summary statistic on level and shape. The three last columns report the *p*-values of the three statistics of each year and the corresponding statistics over all ten years. The last line reports the multi-period test statistic, the sum of the yearly statistics. The null hypothesis of a level and shape calibrated rating methodology is clearly rejected for S&P ratings with a *p*-value of 0.06%, but clearly not for our default probabilities obtained with **MZ1**.